Feuille de TP n°8 – Estimation d'une matrice de transition

On trouvera dans [DCD83] tous les résultats théoriques utilisés dans ce TP.

Pour des applications des chaînes de Markov en biologie, on pourra consulter [RRS03]. Les séquences ADN sont modélisées par des chaînes de Markov dont on veut estimer les transitions, décrire le comportement en temps en grand etc.

1 Estimation de la matrice de transition d'une chaîne de Markov

On considère la chaîne de Markov sur $E = \{1, 2, 3, 4\}$ associée à la matrice de transition

$$P = \begin{pmatrix} 0.2 & 0.3 & 0.3 & 0.2 \\ 0.1 & 0.5 & 0.2 & 0.2 \\ 0.5 & 0.1 & 0.3 & 0.1 \\ 0.3 & 0.1 & 0.2 & 0.4 \end{pmatrix}$$

Notons μ la mesure invariante de la chaîne. On définit, pour $(i, j) \in E^2$,

$$N_n^i = \sum_{p=0}^{n-1} \mathbf{1}_{(X_p=i)} \quad \text{et} \quad N_n^{ij} = \sum_{p=0}^{n-1} \mathbf{1}_{(X_p=i, X_{p+1}=j)}.$$

►► Que vaut la[1] mesure invariante μ ?

Theorème 1. *Pour tout $x \in E$ et tout $(i, j) \in E^2$,*

$$\frac{1}{n} N_n^i \xrightarrow[n\to\infty]{\mathbb{P}_x-p.s.} \mu(i), \quad \frac{1}{n} N_n^{ij} \xrightarrow[n\to\infty]{\mathbb{P}_x-p.s.} \mu(i)P(i, j) \ ;$$

►► Comment estimer la mesure invariante μ à partir de l'observation d'une trajectoire $(X_0, \ldots, X_n)$?

►► Si $(Y_n)_{n\in\mathbb{N}}$ est une suite de variables aléatoires i.i.d. de loi $\nu = (0.2, 0.3, 0.3, 0.2)$ sur E, quel est le comportement de

$$D_n = \sum_{k=1}^{4} \frac{(N_n^i - n\nu(i))^2}{n\nu(i)} \ ?$$

►► À partir d'un échantillon de taille p de même loi que D_n, représenter sur un même graphique, la fonction de répartition F de la loi limite de D_n, son estimation par la fonction de répartition empirique F_p de l'échantillon et une région de confiance pour F déduite du théorème de Kolmogorov-Smirnov.

►► Si tout se passe comme dans le cas où $(X_i)_i$ sont i.i.d. de loi μ, quel est le comportement asymptotique de

$$D_n = \sum_{k=1}^{4} \frac{(N_n^i - n\mu(i))^2}{n\mu(i)} \ ?$$

[1] Pourquoi existe-t-elle ? Pourquoi est-elle unique ?

Que dit la simulation ? Interprétation ? Peut-on tout de même utiliser cette statistique pour tester si la mesure uniforme sur E est la mesure invariante de la chaîne ?
►► Mettre en place ce test. Tracer, en fonction de n, la probabilité (ou au moins son estimation) de rejeter H_0.
►► Proposer un estimateur $\hat{P}$ pour la matrice de transition P construit à partir de l'observation d'une trajectoire $(X_0, \dots, X_n)$.

Theorème 2. *Soit une chaîne de Markov de transition P sur un espace E à s éléments, qui forme une seule classe de récurrence, $\Delta = \{(i,j),\ P(i,j) > 0\}$ et k le cardinal de Δ. On a, pour tout $x \in E$,*

$$D_n = \sum_{(i,j),\ P(i,j)>0} \frac{(N_n^{ij} - P(i,j)N_n^i)^2}{P(i,j)N_n^i} = \sum_{(i,j),\ P(i,j)>0} \frac{N_n^i(\widehat{P}_n(i,j) - P(i,j))^2}{P(i,j)} \xrightarrow[n\to\infty]{\mathcal{L}(\mathbb{P}_x)} \chi^2(k-s).$$

Remarque 3. Le nombre de degrés de liberté peut s'interpréter de la façon suivante : il y a k paramètres non nuls, avec $k \geq s$ puisqu'au moins un coefficient sur chaque ligne de P est non nul, qui sont liés par s relations qui traduisent le fait que P est stochastique.

►► À quoi peut servir ce théorème ? Illustrer l'influence du nombre de coefficients nuls dans P.

2 Pour aller plus loin : un test de markovianité

Pour répondre à la question *la trajectoire observée provient-elle d'une chaîne de Markov ?* sans présupposée connue la matrice de transition, on peut mettre en place un test d'adéquation de loi de $(X_n)_{n\in\mathbb{N}}$ à la famille des lois de probabilités sur $\mathcal{A}^{\mathbb{N}}$ qui vérifient

$$\forall n \in \mathbb{N}, \quad \mathbb{P}(X_{n+2} = k, X_{n+1} = j | X_n = i) = \mathbb{P}(X_{n+2} = k | X_{n+1} = j)\mathbb{P}(X_{n+1} = j | X_n = i).$$

Pour que la machinerie fonctionne, il faut restreindre l'hypothèse H_0 aux chaînes de Markov dont la matrice de transition est à coefficients strictement positifs.

Theorème 4. *Soit $(X_l)_{l\geq 1}$ une chaîne de Markov récurrente sur E fini de cardinal s et de matrice de transition strictement positive. On note, pour i, j et k dans E,*

$$N_l^i = \sum_{n=1}^{l} \mathbf{1}_{\{X_n=i\}}, \quad N_l^{ij} = \sum_{n=1}^{l-1} \mathbf{1}_{\{X_n=i, X_{n+1}=j\}} \quad et \quad N_l^{ijk} = \sum_{n=1}^{l-2} \mathbf{1}_{\{X_n=i, X_{n+1}=j, X_{n+2}=k\}}.$$

Alors

$$Z_l = \sum_{(i,j,k)\in E^3} \frac{(N_l^{ijk} - N_l^{ij}N_l^{jk}/N_l^j)^2}{N_l^{ij}N_l^{jk}/N_l^j} \xrightarrow[l\to\infty]{\mathcal{L}} \chi^2(s^2 - s).$$

Exemple 5. Sur une séquence de 1000 nucléotides que l'on a regroupés en deux classes, 1 pour les purines (c et g) et 2 pour les pyrimides (a et t), on a relevé les résultats suivants :

$$N^1 = 527, \quad N^2 = 473,$$

$$N^{11} = 241, \quad N^{12} = 286, \quad N^{21} = 285, \quad N^{22} = 187,$$
$$N^{111} = 115, \quad N^{112} = 126, \quad N^{121} = 172, \quad N^{122} = 113$$
$$N^{211} = 126, \quad N^{212} = 159, \quad N^{221} = 113 \quad \text{et} \quad N^{222} = 74.$$

►► Quels seraient les estimations des paramètres du modèle de Bernoulli[2] ?
►► Quels seraient les estimations des paramètres du modèle de chaîne de Markov ?
►► Ce modèle semble-t-il approprié ?

Références

[DCD83] D. Dacunha-Castelle et M. Duflo – *Probabilités et statistiques. Tome 2*, Masson, Paris, 1983, Problèmes à temps mobile.

[RRS03] S. Robin, F. Rodolphe et S. Schbath – *Adn, mots et modèles*, Belin, Paris, 2003.

[2]On modélise la suite par des v.a. i.i.d. de loi ν.